様式 C-19

科学研究費補助金研究成果報告書

平成 21 年　5 月 15 日現在

研究種目：若手研究(B)
研究期間：2007～2008
課題番号：19700053
研究課題名（和文）ソフトウエアの動作推定に基づく情報改竄に頑健な不正アクセス検知方式の構築
研究課題名（英文）Robust Detecting Malicious Accesses to Information Manipulation Based on Behavior Estimation of Software

研究代表者
和泉勇治（WAIZUMI YUJI）
東北大学・大学院情報科学研究科・講師
研究者番号：90333872

研究成果の概要：
情報流出事故を引き起こす不正アクセスを検知するために，ソフトウエアの挙動をネットワークトラヒックから推定し，プロトコルヘッダなどが改竄された場合でも適切に動作可能なトラヒックの解析・評価方式を構築した．

交付額

（金額単位：円）

	直接経費	間接経費	合　計
2007 年度	1,600,000	0	1,600,000
2008 年度	1,000,000	300,000	1,300,000
年度			
年度			
年度			
総　計	2,600,000	300,000	2,900,000

研究分野：ネットワークセキュリティ
科研費の分科・細目：情報学・計算機システム・ネットワーク
キーワード：不正アクセス検知，情報流出

１．研究開始当初の背景

サーバを利用しない効率的なデータ通信を可能にした P2P ネットワークが盛んに利用されるようになり，個人対個人での簡便な情報共有が実現している．しかし，P2P ネットワーク上で蔓延する不正アクセスによる情報流出などが新しい社会問題となっている．2005 年に起きたネットワークを経由した情報漏洩は，報道されているだけでも約 130 件程度であるとの報告もある．同報告書では，ネットワーク以外の流出経路を含めた場合の情報漏洩に対する損害賠償額が 7000 億円を超えると試算している．情報流出は，経済的な被害だけではなく，個人のプライベートに関わる問題でもあり，速やかに取り組むべき問題であると言える．

ネットワークを経由する情報流出としては，メールソフトの設定ミスなどの原因もあるが，近年では，P2P ソフトウエアの不適切な利用が原因となっている事故が増えている．P2P ソフトウエアによる情報流出の場合，情報が外部へ洩れたことを利用者が気付かない場合があり，更に問題が深刻となり得る．P2P ソフトウエアによる情報流出を防ぐには，そのトラヒックを検知し遮断する必要があるが，P2P ネットワークを構築するソフトウエアでは，通信に利用するポート番号を改竄することが可能であるため，ポート番号による P2P アプリケーションの特定は不可能である．また，それらのソフトウエアのプロトコルが未公開であったり，暗号通信の利用などによって，効果的な対策が更に困難とな

っている．そのため，ポート番号などの情報改竄に頑健な P2P ソフトウエアの不正利用を検知する方式の確立が急務であると言える．

２．研究の目的

P2P ソフトウエアの不正利用を検知するために，パケットヘッダやヘッダ以外の情報の何が、どの程度必要であり、検知に利用するパケットヘッダの情報をどの程度まで絞り込めるのかを明らかにする。具体的には、ポート番号の情報が虚偽のものであることを前提として、アプリケーションで送受信されるデータの構造やパケット単位での変化を定量的に評価することにより、不正に利用されているアプリケーションの存在やその種別、アプリケーションが異常な挙動を示しているのか否かを同定する方式を提案する。また、トランスポートレイヤーよりも下位のレイヤーのヘッダ情報の利用や、パケットヘッダを全く利用しない場合において、何を観測することによりトラヒックを発生させている不正アクセスや不正利用の検知が可能となるかを明らかにする。

アプリケーションで送受信されているデータ構造の変化に関しては、そのアプリケーションが稼動するために、端末間でやりとりされるメッセージの規則性を抽出し、それを利用したアプリケーション同定方式を確立し、その性能について明らかにする。インターネットで利用されているアプリケーションの一部には、プロトコルが公開されていないものがある。しかし、この場合においても、暗号化通信の鍵交換などの準備段階や、通信相手の稼動ソフトウエアの識別、バージョン情報などの確認のために、一定の規則の則ったメッセージがやりとりされると推定できる。これらのメッセージの規則性を抽出し、そのモデル化を行うことにより、メッセージを発生させているアプリケーションを同定する手法を開発する。

３．研究の方法

(1) アプリケーション毎のトラヒックデータベースの作成
トラヒックダンプデータから、Web やメールなどのアプリケーション毎のパケットを抽出し、送受信端末の組み合わせで定義されるフロー毎のパケットの集合をデータベースに記録する。

(2) 実験ネットワークの構築
本研究の検知対象の一つとして、パケットヘッダに虚偽の情報を設定し不正を生じさせるアプリケーションがある。そのようなアプリケーションを稼動させるための実験ネットワークの構築を行う。想定しているアプリケーションは、P2P アプリケーションである。これらのアプリケーションは、情報流出事故の原因や著作物の流通の中継として利用者の意図とは無関係に利用される可能性があるため、実運用ネットワークとは完全に独立したネットワークでの実験が必要になる。その実験のためのネットワークを構築する。

(3) 特徴量抽出
ネットワークトラヒックを End - to-End のフローに分割し，そのフローから検知に有用な数値情報を抽出する数値化方式を検討する．

(4) 識別方式検討
数値化された情報を可視化し、検知する識別方式の検討を行う．

４．研究成果

研究の方法(3)で検討した数値化方式として，通信に利用されるパケットのデータサイズの逆数をベクトル化する方式と、通信データを構成するコードの発生確率を利用した数値化方式を提案した．提案方式は、各ネットワークアプリケーションには、ユーザー認証、バージョン情報交換などの共通の情報交換が通信の初期段階で行われ、それらの規則性を適切にモデル化出来れば、ポート番号に依存しないアプリケーション識別が可能であるとの仮定に基づいている。

パケットサイズの遷移パターンを利用したアプリケーション識別においては、識別性能を確保するために、一定数以上のパケットを利用することが必要であるが、一定数以上のパケットの通信が行われると、パケットサイズが MTU 付近で一定になってしまい、アプリケーション間の差異が埋没してしまう問題が明らかとなった。そこで本研究では、パケットサイズの逆数を利用することでこの問題を解決した．

パケットサイズを利用した識別では、不要データのパディングなどによってサイズを変更されてしまう問題もあるため、通信内容を利用した数値化手法の提案も行った。この手法は、ペイロードを８ビット毎のコードに分割し、その発生確率を利用してペイロードの数値化を行っている。この数値化方式は、通信内容の不可逆な変換で通信の秘匿性を保持した方式である。フローを構成する各パケットから抽出した８ビットコードの発生確率を２５６次元のベクトルと見做し、このベクトル間の距離の遷移をフローの特徴量とした．

下図は、提案した数値化方式を自己組織化マップ(SOM)で可視化したものである。解析対象として情報流出事故を起こしている Winny

と Share を採用した。図から分かるように、これら二つのアプリケーションが異なる位置に分布していることが分かり、提案した数値化方式がアプリケーションの識別に有効であると言える。

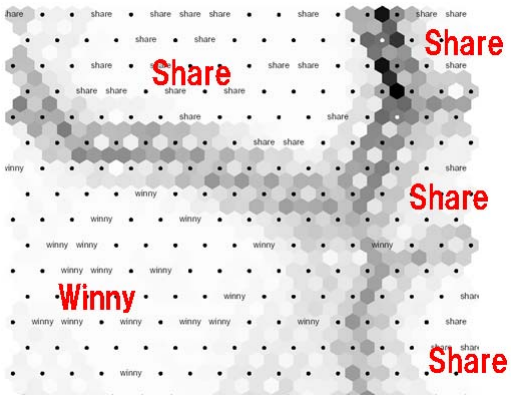

図１：SOM によるトラヒックの可視化

研究の方法(3)で提案した数値化方式を利用し、実際に P2P ソフトのトラヒックを検知するための識別方式を構築した。

トラヒックの可視化に用いた自己組織化マップと同様にベクトル量子化を行うニューラルネットワークのモデルである LVQ(Learning Vector Quantization)を利用し識別方式を構築した。

具体的は、アプリケーション毎の学習データを用意し、それらの数値化と分布の学習を LVQ で行い、未学習のデータの識別を行う方式である。

表１と表２に、それぞれ、パケットサイズの遷移とコードの発生確率を利用した数値化方式による識別結果を示す。Winny と Share 共に、85～98%程度に高い確率で識別出来ている．誤識別の原因として http のトラヒックが非常に広い範囲に分布し、Winny のトラヒックの一部が http と判断されてしまったことが挙げられる。これは、http はプロトコル上、定型のメッセージが短く、バリエーションの大きいＵＲＩや画像、音声、動画などのコンテンツの送受信がトラヒックのほとんどの部分を占めているからと考えられる。http はアプリケーション層のプロトコルであるが、近年では、http 上で動画などのマルチメディアコンテンツの利用が盛んになるなど、http の上位で更に別のアプリケーションが利用されているような状態にある。これは、アプリケーションを識別し、P2P ソフトウエアの不正利用を防止する上では非常に大きな問題となり、今後の研究課題であると言える。

表１：パケットサイズを利用した識別結果

	学習データ	テストデータ
アプリケーション	正解率	正解率
winny	93.10%	85.71%
share	99.31%	98.70%
http	99.72%	99.84%
pop3	96.20%	97.31%
netbios	79.17%	88.89%
skype	72.22%	78.26%
ntp	100.00%	100.00%
dns	40.00%	100.00%
TOTAL	99.05%	99.15 %

表２：コードの発生確率を利用した識別結果

	学習データ	テストデータ
アプリケーション	正解率	正解率
winny	93.10%	85.71%
share	99.31%	98.70%
http	99.72%	99.84%
pop3	96.20%	97.31%
netbios	79.17%	88.89%
skype	72.22%	78.26%
ntp	100.00%	100.00%
dns	40.00%	100.00%
TOTAL	99.05%	99.15 %

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計６件）

① K.Simkhada, T.Taleb, Y.Waizumi, A.Jamalipour, Y.Nemoto, Combating against internet worms in large-scale networks: an autonomic signature-based solution, SECURITY AND COMMUNICATION NETWORKS, Vol.2, 11-28, 2009, 査読有

② S.Yagi, Y.Waizumi, H.Tsunoda, A. Jamalipour, N.Kato, Y.Nemoto, Network Application Identification Using Transition Pattern of Payload Length, Proc. of ＷＣＮＣ Globecom, CDROM, 2008, 査読有

③ S.Yagi, Y.Waizumi, H.Tsunoda, Y. Nemoto, A Reliable Network Application Identification Based on Transition Pattern of Payload Length, Proc. of Globecom, CDROM, 2008, 査読有

④ Yuji Waizumi, Abbas Jamalipour, and Yoshiaki Nemoto, Network Application Identiffication Based on Transition Pattern of Packets, Proc. of IEEE Wi

reless Rural and Emergency Communications Conference, CDROM, 2007, 査読有

⑤ 和泉勇治, 廣瀬淳一, 角田 裕, 根元義章, 相関係数発生確率行列を利用したネットワーク状態評価方式, 電子情報通信学会論文誌 B, 90-B, No.7, 660- 669, 2007, 査読有

⑥ Y.WAIZUMI, M. TSUJI, H. TSUNODA, N. ANSARI, Y. NEMOTO, Distributed Early Worm Detection Based on Payload Histogram Similarity, Proc. of 2007 IEEE International Conference on Communications, CDROM, 2007, 査読有

〔学会発表〕（計０件）

〔図書〕（計０件）

〔産業財産権〕

○出願状況（計０件）

○取得状況（計０件）

〔その他〕

ホームページ等

６．研究組織

(1)研究代表者

和泉勇治（WAIZUMI YUJI）

東北大学・大学院情報科学研究科・講師

研究者番号：90333872

(2)研究分担者

無し

(3)連携研究者

無し